Journal of Business Ethics (2022) 178:945–959
https://doi.org/10.1007/s10551-022-05055-8

REVIEW PAPER

From Reality to World. A Critical Perspective on AI Fairness

Jean-Marie John-Mathews[1] · **Dominique Cardon**[2] · **Christine Balagué**[1]

Received: 1 October 2020 / Accepted: 25 January 2022 / Published online: 25 February 2022

Abstract
Fairness of Artificial Intelligence (AI) decisions has become a big challenge for governments, companies, and societies. We offer a theoretical contribution to consider AI ethics outside of high-level and top-down approaches, based on the distinction between "reality" and "world" from Luc Boltanski. To do so, we provide a new perspective on the debate on AI fairness and show that criticism of ML unfairness is "realist", in other words, grounded in an already instituted *reality* based on demographic categories produced by institutions. Second, we show that the limits of "realist" fairness corrections lead to the elaboration of "radical responses" to fairness, that is, responses that radically change the format of data. Third, we show that fairness correction is shifting to a "domination regime" that absorbs criticism, and we provide some theoretical and practical avenues for further development in AI ethics. Using an ad hoc *critical space* stabilized by reality tests alongside the algorithm, we build a shared responsibility model which is compatible with the radical response to fairness issues. Finally, this paper shows the fundamental contribution of pragmatic sociology theories, insofar as they afford a social and political perspective on AI ethics by giving an active role to material actors such as database formats on ethical debates. In a context where data are increasingly numerous, granular, and behavioral, it is essential to renew our conception of AI ethics on algorithms in order to establish new models of responsibility for companies that take into account changes in the computing paradigm.

Keywords Fairness · Machine learning · Pragmatic sociology · Big data · Business ethics · Artificial intelligence · Responsibility model

Introduction

Artificial Intelligence (AI) algorithms are increasingly used in a wide variety of sectors such as human resources, banking, health, and legal services for tasks such as job candidate screening (Liem et al., 2018), consumer credit scoring (Hand & Henley, 1997), medical diagnosis (Kononenko, 2001), and judicial sentencing (Kleinberg et al., 2018). While these systems are efficient and significantly impact businesses and organizations (Little, 1970), new challenges of various kinds arise when they are used. Algorithm biases, discrimination, and consequently unfairness have been identified in various AI applications, such as predictive models in justice (Larson et al., 2016), facial recognition (Buolamwini & Gebru, 2018), search engines (Kay et al., 2015), advertising (Sweeney, 2013; Datta et al., 2018), speech recognition (Tatman, 2016), AI for recruitment (Leicht-Deobald et al., 2019), and predictive models in healthcare (Obermeyer et al., 2019). Consequently, ethics of AI have become a big challenge for governments and societies, and most of the world's leading universities have recently created multidisciplinary centers of research on AI that focus primarily on responsible AI. The responsible AI movement has spread to most countries and universities, policy makers, companies, and non-profit organizations (Jobin et al., 2019).

Recently, these responsible initiatives regarding AI have come under criticism. At least 84 ethical guidelines (Mittelstadt et al., 2016) have been drawn up to provide high-level principles—such as fairness, privacy, transparency, etc.—as a basis for the ethical development of AI. AI ethics is governed by principlism (Mittelstadt, 2019), meaning

✉ Jean-Marie John-Mathews
jean-marie.john-mathews@imt-bs.eu

Dominique Cardon
dominique.cardon@sciencespo.fr

Christine Balagué
christine.balague@imt-bs.eu

[1] Université Paris-Saclay, Univ Evry, IMT-BS, LITEM, Evry, 91025 Paris, France

[2] médialab, Sciences Po, 27 rue Saint Guillaume, 75 011 Paris, France

that theoretical moral frameworks are produced deductively from abstract principles and then applied to practices. Recent studies criticize principlism for not sufficiently taking into account the particularities of these algorithms and the context of their development and warn against the side effects that can be induced when top-down measures are too prevalent (Mittelstadt, 2019, Powers & Ganascia, 2020). Another responsible initiative in AI—namely Fair ML—is likewise criticized for its top-down and high-level approach. Fair Machine Learning (Fair ML) is a branch of AI ethics involved in technically building fairness-aware learning algorithms to avoid discrimination against minorities (Mehrabi et al., 2019). Using different high-level definitions of fairness, Fair ML integrates ethical constraints into ML algorithms. However, like principlism, recent papers criticize these "fair" techniques because they do not sufficiently take into account the context in which the algorithm is applied and which produces side effects (John-Mathews, 2022; Fazelpour & Lipton, 2020). For example, methods for correcting fairness may actually accentuate intra-category inequality (e.g., inequalities between females) in favor of inter-category inequality (e.g., inequality between male and female) (Speicher et al., 2018). Other authors claim that "fair" algorithms produced by Fair ML abstract away and miss the socio-technical context" (Selbst et al., 2019). They call for a deeper understanding of "the social" implementation of algorithmic techniques to avoid a limited, technologically deterministic, and expert-driven view of AI ethics (Greene et al., 2019).

While the criticisms of AI ethics focus on their lack of integration of the socio-technical context, we show in this paper that a new trend in fair ML is the endeavor to respond to these criticisms by integrating the context through the decategorization of the input data. In this new configuration, the input data become more granular, behavioral, and voluminous. We show that this recent trend—towards softer and de-categorized input data—appears as a response that claims to better integrate the particularities of the socio-technical world and avoid the side effects of the top-down AI ethics.

How does the softening of data structure become a response to the above-mentioned criticisms of Fair ML and therefore a solution to build fair algorithms? What are the limits to this new trend? More generally, can we propose a theoretical framework to understand the debate around the construction of fair algorithms? What are the practical consequences of this theoretical framework in terms of a corporate responsibility model?

To theorize the debate on Fair ML and draw conclusions in terms of business ethics, we argue that pragmatic sociology affords useful concepts to map the criticisms of Fair ML and its responses. As some authors have shown in the field of bioethics, principled ethics are sometimes too high-level and miss the particularities, complexities, and contingencies of real moral issues (Hoffmaster 2018), hence the need for a pragmatic and bottom-up approach to ethics that can take into account contextual aspects. Other authors in business ethics (Martin & Freeman, 2004) have also called for the use of STS and pragmatism to study technology ethics. According to Martin and Freeman 2004, "where business ethicists are making implicit assumptions in their treatment of technology, STS scholars have been explicitly analyzing technology and society in an attempt to understand how the two interrelate". Pragmatism builds ethics from people's lived experience, perceptions, narratives and interpretations and shows how people in their daily lives mobilize and combine multiple normative expectations. By providing reasonable pluralism (Freeman 1994), pragmatism "captures the broader understanding of technology advocated by those in STS and allows business ethicists to analyze a broader array of dilemmas and decisions" (Martin & Freeman, 2004). Pragmatic ethics therefore only acquire meaning in the everyday practices of normal lives when individuals formulate criticisms and expect them to be intersubjectively recognized by others (Galvez et al., 2020). From a pragmatic tradition, this paper does not start from an already defined ethics (such as ethics of duty, virtue ethics, consequential ethics) to explain and understand observed phenomena. We instead seek to describe how a given socio-technical situation can bring out a certain normative reality (Sect. 6) through the mediation of a set of technical equipments (see last section). In other words, we seek to describe how normative categories can emerge in a particular socio-technical situation (situations of algorithmic unfairness when the format of databases changes).

To give meaning to AI ethics and its recent attempt to absorb contextual elements, we use the theoretical distinction between *world* and *reality* as defined by Boltanski, one of the most prominent researchers of the "pragmatic" school of French sociology. Boltanski's pragmatic sociology has been used in numerous works in business ethics to account for complex ethical situations given the particularity and diversity of the various actors' narratives (Galvez et al., 2020; Mercier-Roy & Mailhot, 2019; Dey & Lehner, 2017; Cloutier & Langley, 2017).

The purpose of this research is to shed new light on the current controversy on AI fairness. First, we provide a theoretical framework to understand how ethics in AI is evolving in a context where data and computation techniques change. Second, this research describes how to build a responsibility model that addresses this change using the Boltanskian concept of reality test (*épreuve de réalité*). Finally, this paper shows the fundamental contribution of pragmatic sociology, insofar as it affords a social and political perspective on AI ethics by giving an active role to material actors such as database formats on ethical debates.

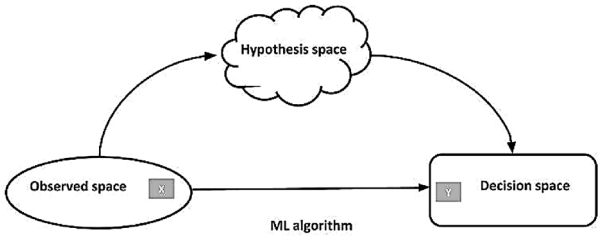


Fig. 1 The three spaces of the ML process

To this end, in Sect. 1 we adopt analogical reasoning to present an analogy between Boltanski's concepts of reality tests and ML algorithmic systems. In Sect. 2, we show that criticism pointing out the unfairness of ML methods takes Boltanski's "realist" point of view because criticisms are based on demographic categories produced by institutions. In Sect. 3, we show that the realist fixing strategy of the Fair ML community is insufficient to address the fairness issue. After presenting, in Sect. 4, the "radical" response to fairness through Boltanski's *world* concept, in Sect. 5 we use Boltanski's framework of managerial domination regime to show that Fair ML methods confuse the *world* with *reality*. Finally, in Sect. 6, we discuss the operational contributions of our theorizing in terms of corporate responsibility model.

Algorithmic Situations as a Reality Test

From Boltanski's Reality Test to Algorithm

We consider algorithmic situations sustained by ML methods as "reality tests" (*épreuves de réalité*) as understood by Boltanski and Thévenot (1983). These reality tests are more or less ritualized moments of social life such as contests, recruitment processes, sports competitions or the realization of a project during which the skills and qualities of people are evaluated and validated. Starting from a state of *reality* constituted by the input data, the algorithmic tests produce choices, rankings or scores that transform the initial situation according to a given objective. To do so, they rely on principles allowing the ML process to select, measure, and help decision-making with relevancy and legitimacy. The operations carried out by ML methods are frequently described in terms of three different spaces (Friedler et al., 2016). The input data *X* constitute the *observed space* and the result of the calculation (*Y* in Fig. 1) makes up the *decision space*. Between the two spaces, the data scientist who sets up the algorithm must imagine a latent space traditionally characterized as the *hypothesis space* in ML process (Mitchell, 1997)—that encloses the ideal variables governing the decision according to the objective[1] (Fig. 1).

For example, when recruiting new employees in a company using an ML algorithm that rates candidates, selected criteria may include the recruitment of intelligent, motivated and competent candidates. But these latent variables—which form the basis for the justification of the algorithmic decision in the *hypothesis space*—do not exist in the actual data from the *observed space X* (i.e., gender, qualifications, past professional experiences, etc.). The latent variables of the *hypothesis space* refer to optimal characteristics on which the final decision to recruit or not will be based. It is therefore necessary to find good approximations in the *observed space* of latent variables in the *hypothesis space*, so that the decision can effectively be made according to the qualities projected in the latent *hypothesis space*. In traditional uses of ML methods, data scientists verify the quality of the articulation between the *observed space* and the latent *hypothesis space* in the feature engineering step (Ghiassi et al., 2016). The idea that a data scientist is able, at least approximately, to produce an intelligible—i.e., "semantic"—interpretation of the link between the *observed space* and the *hypothesis space* has for a long time been the basis of vigilance regarding the quality and fairness of ML models (Mitchell, 1997). For example, the data scientist can check whether, at equal levels of competence (*hypothesis space*), the algorithm creates different decisions (*decision space*) between black and white people (*observed space*).

However, this relationship between *observed space* and latent *hypothesis space* is now increasingly inaccessible for verification and criticism (Burrel, 2016). This is the consequence of the massive transformation in the volume, nature and status of the data used to produce predictions in AI. To highlight the extent of this shift, consider two examples of databases on credit attribution used in the ML literature. The first one is the German credit dataset (Hofmann, 1990). It contains observations (*observed space*) on 20 variables for 1000 past applicants for credit. Each applicant was rated as "good credit" (700 cases) or "bad credit" (300 cases) (*decision space*). Most variables are qualitative such as personal status, sex, credit history, credit purpose, and so on, and can contain multiple encoded categories (*hypothesis space*). For example, the job variable is coded with five possible attributes: unemployed, unskilled, skilled, management, and highly qualified employees. This traditional dataset is characterized by a few number of variables, each one being

[1] The hypothesis space in Machine learning (Mitchell 1997) can be related to Akrich's *script*: engineers and developers of technology inscribe visions or preferences of how the world works (Akrich 1992). Designers of algorithms make assumptions about what the world will do and inscribe how their algorithm will fit into that world, which is very similar to the hypothesis space of Machine Learning engineers.

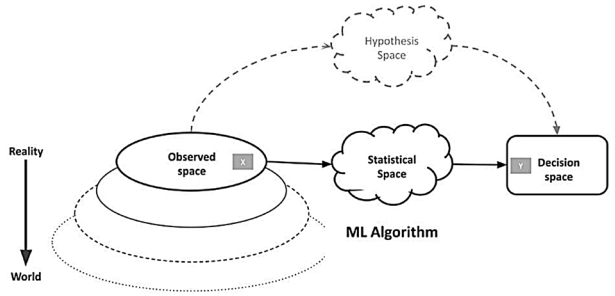


Fig. 2 The shift from *Reality* to *World*

subdivided into a few modalities, providing a stable and homogeneous description of each data point.

The second set of data is a typical example of the new type of information that the digital revolution has made available. Like the former one, it has been used by Lu et al. (2019) used to generate credit scores (*decision space*), in an article that won the Best Paper Award at the International Conference on Information Systems 2019. This dataset is composed of a much larger number of data points and the number of variables has increased dramatically. For each loan applicant, the platform collects personal behaviors like: (i) online shopping records (order time, product name, price, quantity, receiver information); (ii) cell phone-related records (call history, cell phone usage, detailed mobile app usage, GPS mobility trajectories); and (iii) social media usage (whether the borrower has accounts or not, and if so, all posted messages with timestamps and social media presence including the number of fans, followers, received comments, and received "likes" at weibo.com (*observed space*). Each variable is no longer subdivided into a small number of modalities, but into a series of granular pieces of information: shopping list, phone calls, travel, social network behavior. The data are no longer aggregated into variables but appear as the most elementary pieces of information possible. The major difference between both datasets concerns the *observed space*: the first dataset is composed of categorical entities, the second of granular flows of behavioral traces. When the number of input variables increases dramatically to form a vector of several tens of thousands of features (Mackenzie, 2017), it is no longer possible to project a latent *hypothesis space* and to assess the relationship between the *observed space* and the principle that the model is supposed to learn. In this case, the learning method by optimization becomes unintelligible. Without the hypothesis space, it is then very difficult for the data scientist to verify whether the decision of the algorithm is unfair for a certain category of the population (Fig. 2). And yet this shift towards granular flows of behavioral traces concerns more and more fields in Business Analytics, which makes it difficult to address fairness issues in practical cases. This problem is all the more important since ML algorithms are increasingly accused of discrimination. One illustration is the Amazon hiring algorithm in 2015, ranking candidates with a score based on patterns in resumes over a 10-year period, discriminating against women, because the score reflected the male overrepresentation in technical jobs (Dastin, 2018).

Boltanski's Distinction Between Reality and World

The distinction between *world* and *reality* as proposed by Boltanski in *On Critique* (2011) is particularly useful to understand this transformation in the status of data used by recent AI techniques (Cardon et al., 2018). In Table 1, we use analogical reasoning (Hesse, 1966) to map Boltanski's concept of reality tests ("source domain") with ML algorithms ("target domain"). As shown by previous theoretical works in Information Systems (Leclercq-Vandelannoitte & Bertin, 2018), analogy is a central concept for theory building and transfer of meaning from a "source domain" to a "target domain". Each domain is specified by a relational system composed of a set of objects, properties and relations. A mapping function is supposed to exist between the relational systems of each domain (Bartha, 2013).

Table 1 presents analogies linking both domain relational systems. According to Boltanski (2011), *reality* corresponds to a social order that is constantly supported by institutions to produce stable and shared qualifications and representations of the social world. In the recruitment use case for instance, *reality* is about gender, qualifications, past professional experiences, etc. For the credit attribution, *reality* gathers the elements of the German credit system database. *Reality* exists because it is instituted by organized reality

Table 1 Analogy from Boltanski's theory to AI fairness

	Source domain: Boltanski's theory	Target domain: AI fairness
Test	*Reality* tests select and classify social agent	ML algorithm can be used to select and classify individuals
Reality	Statistical categorization produced by institutions	Categories in observed datasets
World	*World* is everything that happens. *World* is unknown, uncertain, and is only represented by *reality* through reality tests	The underlying phenomenon generating data
Criticism	Realist criticism of reality tests is using institutionalized categories	Fairness debate in ML uses institutionalized categories to criticize algorithms

tests of the social world (competitions, recruitment, ranking, qualification, etc.) that are supported by institutions or quasi-institutions. These reality tests define the categories and symbolic representations allowing people to say that a result is fair or unfair (Boltanski & Thévenot, 1983). It is also on the basis of these same categorical representations that the results of reality tests can be criticized—and we will call this "realist criticism". By using other categories (such as sex, gender, qualifications, etc.), it is possible to show that the results of a reality test, such as contests for enrollment at academic institutions, are unfair because they produce discrimination by favoring children from more privileged backgrounds. However, results of reality tests can be contested from a broader perspective that reveals the artificiality and the arbitrariness of categories serving to stabilize *reality* and our ordinary sense of justice. Boltanski (2011) defines *world*—"everything that happens"—as the background against which it is possible to denounce the artificiality of *reality*. The test then appears unfair because it does not take into account the singularity of the candidates' profiles. This singularity is not captured by the existing categories because the test criteria are imperfect, since other dimensions of the *world* should be included in the computation. The approach of pragmatic sociology considers that we should think of the social world as a set of ongoing processes rather than as a collection of social entities and structures (Abbott, 2016). Criticism of the results of the reality tests can then be called "radical"—or "existential" (Boltanski, 2011, p. 103)—in the sense that it emphasizes the existence of a denser and richer world of facts than the information used in the selection process. As our second example with Chinese loan application data shows (Lu et al., 2019), big data claim to record a broadening and singularization of information. We think that this distinction between *reality* and *world* allows for a theoretical reading of an essential shift that is taking place today in the prediction techniques of AI. Until recently, statistical methods have been used extensively, based on categorical data produced by organizations (states, companies) that can be understood as *reality*, as defined by Boltanski (2011). However, with the development of huge numerical databases, current calculators can produce predictions by integrating more and more elements of the *world* and by considerably increasing the number of data, variables and information on which calculated predictions can be based.

Realist Criticism on Machine Learning Unfairness

Many cases of ML unfairness have been identified and denounced by academics and the media in the last few years. We argue in this paper that these criticisms belong to a "realist" vision, on two grounds: (1) criticism is mostly supported by two "protected categories", namely race and gender; and (2) they are using statistical indicators of fairness calculating differential treatment by the ML algorithm, between these "protected" categories. This criticism is therefore realist since it is based on categorical and stabilized representations of reality by proposing to redress/correct the relative share of one category in relation to another. The implicit theory of justice that it implements thus needs to rely on a system of categorization whose meanings are shared and validated by institutions in order to correct the harms of unequal distribution.

Criticism of Machine Learning Unfairness Anchored in Boltanski's Reality Vision

The ordinary meaning of just and unjust is based on the representations that individuals have forged in a social world, and which in turn have shaped categories produced by institutions and companies (Boltanski & Thévenot, 1983; Desrosières, 1998). The demographic representation of the distribution of certain categories informs types of criticism that evolve according to historical conjunctures. In the academic literature denouncing algorithmic unfairness, criticism mainly focuses on sensitive categories, such as gender or race. In the use case of ML algorithms for recruitment, discrimination against women has been shown for stem jobs (Lambrecht and Trucker 2019; Dastin, 2018), while it constitutes an infringement of the law. Moreover, other criteria of discrimination constitute potential discrimination issues in HR ML algorithms for recruitment, such as age, race or disability. In what concerns the ML algorithms for credit attribution use cases, regulation requires to respect non-discrimination on categories such as race, gender, address, origin or religion. More globally, in order to denounce the existence of a bias in algorithmic decisions, it is therefore necessary to compare the representation of a category in the *observed space* and in the *decision space*. This is mainly done by fairness metrics that compute demographic distribution in *decision space* between categories in *observed space*.

Production of Fairness Metrics

A recent community of ML researchers is trying to propose technical solutions to mitigate those kinds of algorithmic biases and to address discrimination. This community—often referred to as Fair ML—is increasingly popular in ML. Many international conferences have addressed the subject, such as ACM FAT, FairWare or AIES. Fair ML defines fairness indicators formally, to correct biased ML algorithms. Table 2 shows different fairness metrics put forward by the Fair ML community (Haas, 2019).

Table 2 Fairness metrics

Fairness metrics	Definition
Statistical parity	Probability of being classified with the favorable label is independent of group membership
Disparate impact	Ratio of probabilities of being classified with the favorable label between protected and unprotected groups is close to one
Equalized odds	Both false positive rates and true positive rates for protected and unprotected groups are the same
Equal opportunity	True positive rate is the same between protected and unprotected groups
Predictive rate parity	Fraction of correct positive predictions is the same for protected and unprotected groups

Each of these fairness metrics is defined as demographic differences between two groups in a symbolic and stabilized way. Criticism of algorithmic fairness using these metrics is then grounded in Boltanski's *reality* since they are supported by stabilized categorical representations. For instance, in ML algorithms for recruitment, one fair ML method to limit race discrimination consists in adding a constraint to the algorithm integrating equalized odds, with the same false positive/true positive rates for black and white individuals. If we apply the equal opportunity fair ML method in algorithms for credit attribution, the data scientist will force the system to have the same true positive rate for women and men.

Repairing Injustice by Fixing Algorithms

The research conducted within the Fair ML community mainly aims at finding solutions to integrate corrective measures in the calculation of ML methods. To do so, the community embeds the above fairness metrics into the construction of the algorithm itself. As a result, these methods address and absorb the liberal criticism of discrimination by building ML algorithms that optimize a chosen fairness metric by design (Hoffmann 2019). We define realist fairness engineering as the Fair ML techniques, based on categories, used to address algorithmic fairness. These are realist answers to criticism, in Boltanski's terms, since they rely on a representation between ideal variables, like skills, in a latent *hypothesis space* to assess and correct algorithmic fairness. Actually, they use categorical fairness metrics to correct ML algorithms without changing data formats and models.

Realist Fairness Engineering

Realist fairness engineering consists of two steps. First, data scientists select a fairness metric adapted to the context of implementation and use of the algorithm. Since fairness metrics can be contradictory and antagonistic to each other, the context of use of the algorithm must be specified in order to associate the appropriate fairness metric with it (Kleinberg et al., 2016). Then, in a second step, the algorithm that best verifies the fairness metric is built mainly by means of three different techniques that are commonly listed in the Fair ML literature: pre-processing, in-processing, and post-processing (Mehrabi et al., 2019). Pre-processing techniques aim at modifying the training dataset so that solutions with greater fairness are calculated (Zemel et al., 2013). In-processing techniques adjust algorithms themselves during the learning process so that the resulting classifier has a high degree of fairness (Kamiran & Calders, 2010). These methods incorporate fairness metrics as a constraint in the learning process. Finally, post-processing techniques aim to achieve fairness by adjusting the solution provided by the classification algorithm. This can be done by changing decision thresholds or the decision itself for some data points (Hardt et al., 2016).

A practical application of this philosophy is the IBM fairness 360 toolkit (Bellamy et al., 2018). This toolkit provides several learning methods (pre-learning, in-learning, and post-learning) and different fairness indicators (statistical parity, equal opportunity difference, disparate impact) based on categories to address fairness. For each learning method, the toolkit quantifies the improvement of fairness with respect to the different fairness indicators. In order to select the right algorithm, the user is free to test the different learning methods and choose the most effective one according to the fairness indicator he or she has previously chosen.

Limitations of Realist Fairness Engineering

The integration of fairness corrective measures within the calculation is not enough to address criticism of algorithmic unfairness. A prominent criticism in today's debate is that biases are rooted in the structures of the social world. Fair ML is proposing methods to protect "sensitive" groups from biases. These groups can be women, black people, gay people, minorities and so on. There is already a great deal of research and work in social science on gender equality and anti-racism. These studies often assume the existence of structural power relations between groups, like patriarchy or institutional racism, that are grounded in historical and social constructions. If we suppose this world structure

reflects *reality*, Fair ML papers are not addressing world issues but only data available to them. These data are moreover human constructions, so they incorporate power relations and therefore need to be managed as social constructions. In that sense, data integration, transformation, and modeling are operations rooted in a social world and are therefore contaminated by power relations within the world (Crawford & Calo, 2016; Eubanks, 2017). As an example, we can imagine a team of data scientists working to make hiring algorithms less discriminatory towards women. But since all the practices of these data scientists are rooted in a socio-technical world that already encloses relations of domination, this data scientist team can still have discriminatory practices. For example, this team of data scientists can be mostly composed of men while fighting against discriminatory algorithms. In that case, fairness fixing is not enough since biases are structural.

Statistical correction of biases, referred to as "realist answers" to criticism of unequal distribution of resources among categories, in Boltanski's terms, leaves many questions open. We use concepts from pragmatic sociology and its criticism of categorical approaches to propose some limitations of realist fairness engineering. Luc Boltanski's notion of world refers to a set of traits, affects, relationships to others and things that are not taken into account in objectification techniques conventionally used in institutional reality tests. The *world* constantly overflows *reality* to challenge codification techniques, categorization and decision-making rules of the devices/social systems that support it.

Representation Can Be Too Simplistic

When the *hypothesis space* is too simplistic, it may fail to represent the singularities of the social world. The social world as seen by pragmatic sociology is not composed of fixed entities that can be described by variables that have monotonic causal relationships with each other. In opposition to the idea of a "general linear reality", Andrew Abott's processual sociology assumes that the meaning of entities changes according to place, situation or interaction (Abbott, 1988). Hoffmann (2019) shows that antidiscrimination discourses used with single-axis thinking (i.e., gender or race) are limited because they fail to address social and contextual issues. Some recent works show the limits of Fair ML methods and the potential adverse effects of their implementation (Fazelpour & Lipton, 2020). Selbst et al. (2019) even show that these "Fair" methods are ineffective, inaccurate, and sometimes dangerously misguided when they enter the societal context that surrounds decision-making systems. For example, the value of a given fairness metric (e.g., demographic parity) is always inaccurate, since it is based on historical data testing environments, which can highly differ from the socio-technical world.

Intra- and Inter-category Fairness

The categories used to measure fairness produce groups that are too large and too homogeneous by ignoring intra-category variability. These representations of the social world are being increasingly criticized for their normalizing effect and their inability to take into account subjective dimensions of the sense of injustice or discrimination. Methods for correcting fairness may involve accentuating intra-category inequality in favor of inter-category equality. Correcting unfairness by changing decisions made by the algorithm for an individual in the protected group may consist in favoring those who are most privileged in this protected group. In the ML algorithm for recruitment, one well-known example is positive discrimination, favoring privileged individuals in the protected group, and creating intra-category inequality. These methods can "lead to systematic neglect of some injustices and to misguided mitigation strategies" (Fazelpour & Lipton, 2020). Methods based on broad and rigid categories are too simplistic and do not consider the intersectionality and dependencies between attributes of variables (Buolamwini & Gebru, 2018).

Trade-off Between Performance and Fairness

When fairness is measured by a fairness metric, there might be a trade-off between algorithm performance and fairness (Chatterjee et al., 2009). As an illustration of this trade-off, ML algorithms for recruitment can have place of residence as an input. Since place of residence is generally correlated with race, removing this variable may increase the fairness of the algorithm (with regard to the issue of racial discrimination). However, place of residence can also be correlated with skills and removing this variable may degrade the performance of the algorithm. Since the interest of algorithm designers may be on performance rather than fairness, voluntary fairness correction is hard to expect without a binding mechanism.

A Radical Response: When Algorithms Compute the *World*

As we have just seen, reality tests built from categorized databases are very sensitive to criticism of the reductive, imprecise and arbitrary nature of the variables used to measure injustice and discrimination. They provide a basis for only a reformist and realist approach to bias mitigation. These aporias can explain the very profound shift taking place today regarding methods in AI. A new paradigm is emerging by transforming: (1) the nature of the data; (2)

the methods of calculation; and (3) the type of domination it exerts over society by making criticism of bias almost impossible.

From Observed Data to World: Construction of Fair Representation by Constantly Adding New Data "from the World"

Emerging promises associated with big data and AI are driving a major shift in the format of data. Facing criticism that traditional statistical categories misrepresent *reality*, the new algorithmic tests seek to reduce the arbitrariness and imprecision of reality tests by capturing the *world*. This shift is characterized by a considerable increase in the number of data, the granularization of statistical entities (which become sub-symbolic and can no longer be represented by categories referring to an interpretable type), the personalization of information, and the accumulation of behavioral traces. For example, AI opens personalized recruitment processes, by analyzing candidates' reactions to interactive information on the company in job interview videos or even using external data, such as purchasing behavior or other traces in the social media (Sánchez-Monedero et al., 2020). These recruitment methods that use behavioral traces are supposed to be more accurate and performant than recruitments based on socio-demographic categories or candidates' degree, according to the promoters of these methods.

There is a recent call in the Fair ML community to find more precise and voluminous data about the target variable, to address algorithmic unfairness. Lu et al. (2019) show that using additional alternative data (cell phone usage and mobility trace information) in financial credit risk assessment improves fairness towards lower-income and less-educated loan applicants. In this paper, no causal structure of credit default in a latent *hypothesis space* is needed to improve fairness. Adding fine-grained user behavior data with cell phone usage and mobility trace has been shown to suffice for disadvantaged populations to be included in the credit offer. Similarly, the paper of Cai et al. (2020) claims that disparities across groups come from different qualities of information across groups. They suggest that decision makers should "…screen candidates on the margin, for whom the additional information could plausibly alter the allocation…" Cai et al., (2020, p. 1). In that sense, achieving fairness requires the collection of additional data and "screen off" more candidates to improve the quality of information about individuals.

World Reconstruction as a Statistical Space Inside the Algorithm

This shift in data format also comes along with a change in calculation techniques. If data are more decomposed and granular, model parametrization becomes high-dimensional like that in the Deep Learning method. An ex post latent approximation of Boltanski's *world* is then supposed to be encoded or embedded by the high-dimensional parameters inside the algorithm. The ex ante latent *hypothesis space* from realist responses turns into an ex post latent data representation between the *observed space* and the *decision space* inside the algorithm. We refer to this intermediary latent space as the *statistical space*[2] (see Fig. 2).

How is this *statistical space* created inside the algorithm in ML? Learning low-dimensional representation from high-dimensional and voluminous data is a common practice in ML (Pan & Yang 2009). In the case of algorithms for recruiting, integrating behavioral traces (such as video data from the candidate's job interview or the candidate's buying behavior on the Internet) is expected to automatically reconstruct new latent representations, which were inaccessible in the realist configuration, such as propensity to smile, propensity to have impulsive buying behavior, etc. Some authors similarly suggest methods to learn "Fair representation". For example, Zemel et al. (2013) "formulate fairness as an optimization problem of finding a good representation of the data with two competing goals: to encode the data as well as possible, while simultaneously obfuscating any information about membership in the protected group" (Zemel et al., 2013). The authors define these representations as "Fair representation". Other authors construct these "fair representations" using adversarial networks (Madras et al., 2019). In that case, the aim is to create a neural network designed both to predict the target variable and to wrongly predict the sensitive variable using an adversarial network. This creates a latent *statistical space* that is both agnostic to the sensitive variable and designed to predict at best the target variable. This latent space is interpreted as a constructed fair space. More generally, among methods using an individual fairness definition, there is an assumption that an "ideal feature space exists […], and that those features are recoverable in the available data" (Suresh & Guttag, 2019). For example, Louizos et al. (2015) propose "learning representations that are invariant to certain nuisance or sensitive factors of variation in the data while retaining as much of the remaining information as possible". These representations are fair latent spaces abstracting the sensitive variable (sex, race) to achieve fairness.

While radical responses seem to annihilate the ex ante latent *hypothesis space* of realist methods (by breaking the path between *observed space* and *decision space*), they actually return it ex post in a materialized form of *statistical*

[2] Latent spaces in Machine Learning are linear combinations of the *observed space*. For example, in the case of neural networks, the *statistical space* is the hidden layers of the neural network.

space. In other words, the project of the "radicalized" Fair ML methods could be interpreted as a recomposition of the entities of the world by inscribing or materializing them within the calculation itself (*statistical space*). To do this, these methods extend the calculation to a maximum number of entities in the world.[3] If this project seems to be part of pragmatic sociology (Latour, 2005), we will see in the following section that it is not: it confuses the *world* with *reality* and locks the calculation into what Luc Boltanski calls a complex regime of domination.

Complex Regime of Domination Confuses the *World* with *Reality*

From Simple to Complex Regime of Domination

Reality tests are tools of power. They are designed to prioritize, rank and distribute opportunities within our societies. In this way, they create asymmetries between those who pass the tests and those who fail. For this reason, they must receive some form of legitimacy: members of society must consider the results of the test to be "fair". While algorithmic techniques are becoming increasingly important within reality tests (recruitment, scoring, access to credit, legal decisions, etc.), it is important to analyze how modes of domination are changing in the age of big data. In his analysis, Luc Boltanski contrasts two forms of domination: "simple domination" and "complex or managerial domination" (Boltanski, 2011, p. 133). Whereas simple domination seeks to confirm the legitimacy of reality tests by merely correcting the dysfunctions highlighted by criticism, managerial domination constitutes a new configuration of the exercise of power marked by the ongoing change of reality tests. It consists in "continuously modifying the contours of *reality* as if to inscribe the *world* in it" (Boltanski, 2011, p. 133). In order to deal with criticism that constantly underlines the gap between the principles put forward to justify the tests and *reality*, the new computation paradigm described above focuses on constantly changing the format of the tests.

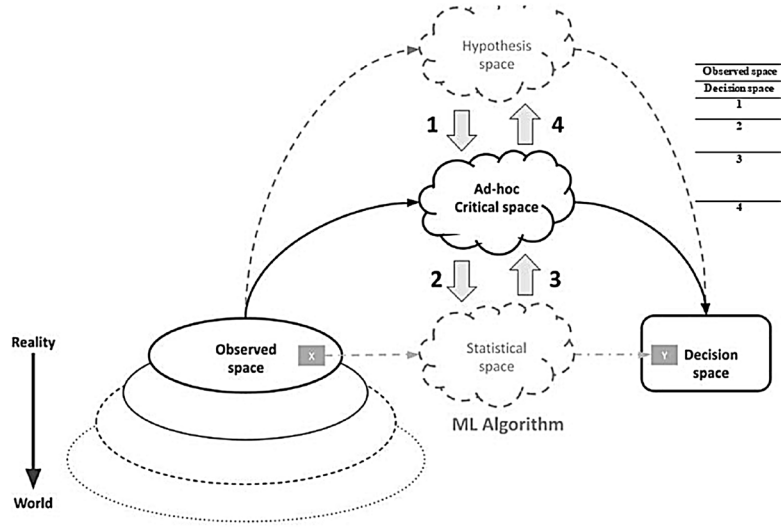
Several characteristics of the complex/managerial domination regime can be highlighted. First, as we have just seen, algorithmic tests now rely on a massive and continuous expansion of the volume and nature of data. Second, the information from the *world* embedded in the calculator has no longer categories that users can understand and from which they can make assumptions about the *world*. They are flows of behavioral traces (telephone interactions, Internet browsing, geolocation, etc.) from which it is very difficult to extract a relevant meaning—which, on the other hand, algorithms do. The *world* thus appears more and more as a snapshot of "life itself" (Rouvroy & Berns, 2013). Third, the implementation of this new type of test depends more and more on scientific expertise. "In the political metaphysics underlying this form of domination, the *world* is precisely what we can presently know, through the powers of Science" (Boltanski, 2011, p. 131). Animated by the positivist and naturalistic project to conduct a physical science of society (Pentland, 2014; Zuboff, 2019), experts are the only ones qualified to access the *world* and make it available to calculators. In this way, they arrogate to themselves the possibility of constantly rearranging reality tests. The complex domination regime is thus called system of experts' domination ("système de domination des experts") by Boltanski. Experts no longer derive their justification from intelligible principles, but from the idea that observable data encompass the *world* as adequately as possible.

Consequences of the Complex Domination Regime

One of the major consequences of this shift from *reality* to the *world* is that since computational data can no longer be organized in the form of interpretable symbolic variables, it is no longer possible to project them into a *hypothesis space*. In other words, the ex post fair representation becomes harder to interpret when models and data are high-dimensional (Molnar, 2018). Even though this ex post representation may encode very fine-tuned dependencies between variables missed by realist categories, there is no chance of guaranteeing fairness claimed by realist criticism without the possibility of interpreting this latent data representation (John-Mathews, 2021). Algorithmic opacity is one of the greatest practical and business challenges posed by AI and is today the subject of many research works (Burrel, 2016). For example, it is important for an algorithm of recruiting to be able to give an explanation of its decision-making process. Explanations are necessary to ensure confidence, prevent errors and address contestation from refused candidates.

Contrary to the radical approach, normalized data categories from the realist approach have the advantage of giving a representation to the data and to the mechanism underlying the algorithm. Since criticism is realist using mainly categorical variables, the radical response using non-interpretable ex post *statistical space* cannot be compatible with realist criticism. In a way, the radical approach moves the representation of the data away from the observed data and encodes it inside the algorithm through a latent space. The challenge is then to equip and open-up this *statistical*

[3] "Reality" is biased because it can constantly be criticized by new normative expectations that arise from the "world". As a result, in Luc Boltanski's conceptual framework, the world ("everything that happens") cannot be biased as such. Biases only exist when viewed from the categories of reality (which alone can reveal the existence of unequal distribution or discrimination).



	Example
Observed space	Candidate's video of the job interview and purchasing behavior data
Decision space	Having the best evaluation two years after recruitment
1	**Break down the objective.** Ex: what does it mean to be a "good candidate"?
2	**Radicalizing through experimentation.** Ex: candidates' purchasing behavior can be artificially modified to experiment algorithmic outputs.
3	**Stabilize new realities by aggregating points of view.** Ex: emergence of new categories like "candidates smiling during the interview but with impulsive buying behaviour"
4	**Re-qualify the objective and open-up realities.** Ex: Conduct a multi-stakeholder discussion to assessif the emergent categories can be normative categories for making a recruitment decision

Fig. 3 Stabilizing emergent normative realities to build a responsibility model

space in order to guarantee fairness and build a responsibility model.

Affected individuals are no longer able to understand, verify and criticize the fairness of algorithms. This exclusion can be detrimental to criticism or negotiation. As Luc Boltanski points out, the new regime of complex domination through which experts constantly change the reality tests by multiplying new variables extracted from the *world* is also a means of preventing criticism. A recent study (Saha et al., 2020) shows that even the definition of the simplest fairness (demographic parity) is hard to understand and to apply, and that people tending to agree with the fairness definition are not likely to understand it. Another study shows that probability formalism has very different understandings depending on the level of education (Zhou & Danks, 2020). Other recent studies emphasize that it is possible to fool ML interpretation methods so that bias is hidden within the algorithm (Bastani & Bayati, 2020). For instance, a candidate who feels discriminated by ML algorithms in a recruitment or a credit attribution process will encounter difficulties in defending his right to a judicial remedy. More generally, the opacity issue could lead to the emergence of a two-tiered society where some are excluded from the capacity to criticize, while others know how to circumvent the rules.

How to Stabilize Emergent Normative Realities to Build a Responsibility Model?

We argue in this paper that the *reality* as it appears in the hypothetical space is essential to build a responsible model for AI (Martin, 2019a). Without a stabilized and meaningful understanding of the roles and causes of AI outputs, it is indeed not possible to engage the responsibility of the AI designer. Even though traditional categories have limitations in identifying all forms of injustice (see Sect. 3), we argue that we should not abandon *reality* but instead stabilize new realities that are radical enough to take into account the change in data format but without falling into the system of complex domination that encloses the calculation. How can this oxymoron be resolved? Pragmatic sociology offers an original approach to answer this question. Science and Technology Studies indeed proposes a radical method for describing the *world* that avoids a naive positivism—confusing the *world* with *reality*—and without falling into the system of domination of experts that encloses any critical debate (Latour, 2005). On the one hand, in order to avoid naive and naturalistic induction (confusing the *world* with *reality*), pragmatic sociology gives materiality to the *world*'s entities by questioning it through active intervention. On the other hand, in order to avoid the domination of experts in the construction of *reality*, pragmatic sociology opens up the description of the *world* by exposing it through a public investigation, confronting it with contradictory discourse, and fostering reflexivity.[4]

In this section, we propose a mechanism of criticism anchored in a changing socio-technical *world*. To do that, we think that we need to build new reality tests *alongside* the algorithm that enlist new actors with materiality and reflexivity, this is the *critical space*. At the interface between

[4] These two aspects from pragmatic sociology (materiality and reflexivity) is closely related to other research in business ethics that put forward embeddedness and reflection for responsibility model in AI (Martin 2019b).

the *hypothesis space* and the *statistical space*, this ad hoc *critical space* should provide a grip on the socio-technical *world* while avoiding the system of expert domination that prevents criticism. It needs to create the conditions of possibility for the emergence of normative realities that take into account the changing socio-technical context. Figure 3 provides a representation of the different operations that can be performed by this *critical space*. The *critical space* is animated by a double movement: radicalization of the *hypothesis space* (steps 1 and 2) and "putting into *reality*" the *statistical space* (steps 3 and 4). The radicalization of the hypothesis space is carried out by exploring the different paths between the *observed space* and the *decision space* opened up by the *statistical space*. Once the different paths are sufficiently deployed, the second movement recomposes the entities described in the previous step by aggregating them according to various points of view. The ad hoc *critical space* materializes or equips this double movement of exploration and recomposition with a set of tools for analysis, data manipulation, simulation, visualization, aggregation, ad hoc surveys, debates, etc. Materializing the path between the *observed space* and the *decision space* is essential to circumvent the system of expert domination that totally abandons the possibility of following the traces of transformation between the *observed space* and the decision-making space.

To illustrate these four operations, we take the example of a recruiting algorithm using a candidate's both video of the job interview and purchasing behavior as input data (*observed space*). We consider in this schematic example that the ML algorithm has to predict the candidates who are expected to have the best evaluation 2 years after recruitment as an employee (*decision space*). The objective of the algorithm is to select the best candidates (*hypothesis space*).

(1) *Break down the objective* the first step of this ethical exploration of the algorithm's operation is to break down the objective of the calculation in order to make designers sensitive to the diversity of paths leading to the result. Given the complexity of the calculations performed by ML techniques when the variables are very numerous, designers can only make hypotheses about the criteria which seem to them the most relevant to reach the objective. What does it mean to be a "good" candidate for a company in a recruitment process? Who defines this objective? What are the most important criteria to define a "good" candidate? Are hard-worker, pleasant, competent possible criteria? These questions allow the definition of the axes of investigation in step 2.

(2) *Radicalizing through experimentation* based on the questions of the previous step, the algorithm can be experimented through the manipulation of input data and the visualization of the output to better understand the algorithm's decision process. For example, using simulation tools, candidates' purchasing behavior can be artificially modified in the *observed space* to observe if the recruitment decision changes in the *decision space*. As the *statistical space* is often complex (hidden layer of neural networks), we consider that experimentation through active interventions of input data is the appropriate way to have a local intuition of the behavior of the algorithm. Each of these manipulations gives a different point of view on the *statistical space*. These points of view will then be used in the next step to stabilize new "realities".

(3) *Stabilize new realities by aggregating points of view* using visualization tools, one shall multiply the experimental interventions from the previous step until stable trends appear, these are the emerging realities. Since the algorithm draws from an infra-reality (the so-called *world*), we must succeed in constructing the categories that are necessary to meet our normative expectations. The creation and stabilization of these realities will certainly require the mediation of tools for visualization and interpretation that could be commonly interpretable. Some prospective studies have recently proposed to create such an ad hoc *critical space* of very deep neural network (DNN) through interactive visualization tools (Olah et al. 2020). For example, after intervening on purchasing behaviors, the algorithm's outputs can be observed depending on the smiling behavior that is seen in the candidates' video during the interview. Suppose we observe that the algorithm overselects candidates who smile during the job interview (over-representation of false-positives among smiling candidates). But let's also suppose that, using simulation tools, we notice that smiling candidates with impulsive purchasing behaviors are actually bad employees after recruitment in the company. These different experimentations of input data create an emerging new category: candidates smiling during the interview but with impulsive buying behavior. This emerging category will then be requalified and debated in step 4.

(4) *Re-qualify the objective and open-up realities* can the category of "candidates smiling during the job interview with impulsive purchasing behavior" be a normative category for making a recruitment decision? Who are these people in terms of socio-demographic variables? Qualifying emerging realities may involve carrying out ad hoc surveys on individuals in the learning base, in order to retrieve new categorized data (socio-demographic variables for example) to construct a realist representation a posteriori. This ad hoc qualification of existing data therefore requires the mediation of techniques, such as survey tools but also governance mechanisms that can involve external auditors,

ethical charters, or ethical committees. Empirical future research in business ethics needs to be conducted to better specify the practical modality of such a realist governance. We don't provide any "right properties" for these emerging normative realities. However, we propose precise criteria for the process of formation of the normative realities. This process, that we call the ad hoc *critical space*, must be open to the socio-(technical) *world*. Therefore, as many people as possible are invited (end users, data scientists, ethicists, AI product managers, etc.) to manipulate and stabilize normative realities according to their own representations.

Limitations and Future Research

The approach developed in this article is mainly speculative. Drawing on the notions of *world* and *reality* from Luc Boltanski's pragmatic sociology, we show how the joint transformation of data structure and algorithmic computation calls for a reassessment of the way we think about fairness or discrimination in debates on AI ethics. First, we show that criticisms of ML unfairness are "realist", that is, they are grounded in an already instituted *reality* based on demographic categories produced by institutions. Second, we show that the limits of realist corrections lead to the elaboration of radical responses to fairness, that is, responses which radically change the data format and the computation technique. By increasing the variety of initial data to capture a richer and denser *world*, new approaches to algorithmic prediction claim to be more authentic and truthful. Thirdly, we show that these new techniques contribute to obscure calculations by preventing them from criticism and by conferring an exorbitant power to experts through "a constant change of data models". However, the objective of this contribution is to show that one should not renounce to submit these calculation techniques to a rigorous ethical evaluation (Ananny & Crawford, 2018). Therefore, our fourth contribution is to propose a responsibility model relying on an ad hoc *critical space* avoiding experts' domination, to build normative realities through multi stakeholders' discussion.

For this, two conclusive reflections should guide future research.

The approach we have developed in this article is part of a set of works criticizing the top-down approaches of the major AI ethical charters for being too vague (Mittelstadt, 2019). These ethical principles fail to capture the diversity of injustices and discriminations that algorithms can reveal when the data that feed their calculations are not institutional categories but multiple, diverse, and overflowing signals of their identity (Fazelpour & Lipton, 2020). These multiple signals (instead of one category such as gender or race) is a key issue in sensible domains such as health, recruitment, justice, finance, where discrimination effects can highly impact people's lives. In contrast to a top-down approach to ethics, the pragmatic approach places particular importance on the fact that norms for a fair distribution of resources must always be appreciated in situations and that they are embedded within the socio-technical infrastructure (Jaton, 2021; Morley et al., 2020). When the format of the data infrastructure changes, the criticism of the algorithms' output is necessary through a multi stakeholders' discussion in order to incorporate new normative expectations. The main future challenge for AI ethics practitioners will be to succeed in testing and evaluating forms of injustice that will no longer be expressed through sensitive categories such as gender, age, degree or race, but through more composite entities, such as, for instance, in recruitment process, candidates smiling during the interview with impulsive buying behavior. As demonstrated by STS's research, it will then be necessary to translate these compound entities into forms that are easier to interpret and manipulate. This pragmatic approach encourages awareness of the reflexive capacity of individuals to appreciate with hindsight the ethical values that must be prioritized according to the materialized context. Pragmatic sociology seems to us particularly relevant to capture the primordial role played by different data formats and computing configurations in how we give meaning to ethics. But if the top-down approach to algorithm ethics is too vague, a pragmatic one might be too flexible (Floridi, 2019). The main future challenge for AI ethics practitioners will be to succeed in testing and evaluating forms of injustice that will no longer be expressed through sensitive categories such as gender, age, degree or race, but through more composite entities such as for example candidates smiling during the interview in the recruitment process with impulsive buying behavior. As demonstrated by STS's research, it will then be necessary to translate these compound entities into forms that are easier to interpret and manipulate. Responsible governance of the decisions made by algorithm designers must be implemented within companies. It must be supported by a clear process of review and evaluation based on a charter of principles, as highlighted by numerous works relating the implementation of ethical principles within organizations (Clarke, 2019). With respect to the pragmatic approach elaborated in this article, making the trajectory of prediction visible and publicly debatable within the ad hoc *critical space* is essential.

An important limitation of the approach developed in this paper is that it hypothesizes that current research in computer science and data visualization will indeed succeed in producing tools to explore the new statistical entities produced by ML methods. Due to the rate at which large database processing is developing today, the development of new methods and studies on the explanation of DNN decision-making has become an active field of research. A

body of research carried out notably in the computer vision community is focused on the development of tools and methods to "see" and explore the workings of deep learning techniques (Olah et al., 2020); other works concern the treatment of large volumes of text by Natural language processing methods (Kim et al., 2020). They hypothesize that even if the explicability of calculations is unachievable, it is nevertheless possible to understand the paths taken by algorithmic predictions within datasets. These highly promising approaches remain experimental for the moment. They retain a great complexity that complicates the explicitation of the principles of computation and, therefore, an ethical examination. We believe, however, that future research should develop an interdisciplinary approach involving specialists in neural networks, STS, and ethics. The research program that such a collaboration opens up is to succeed in concretely designing what we have called in this article an ad hoc *critical space* that enables to stabilize the computational flows of new algorithms through reality tests. We are currently building and conducting experimental research on such a *critical space* with our experimental project called *AI Autopsy Platform* that allows the dissection, exploration, and diagnosis of AI following ethical incidents.

Funding This research received support from Good In Tech Chair, under the aegis of the Fondation du Risque in partnership with Institut Mines-Télécom and Sciences Po. No funding was received to assist with the preparation of this manuscript. No funding was received for conducting this study. No funds, grants, or other support was received.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

References

Abbott, A. (1988). Transcending general linear reality. *Sociological Theory, 6*(2), 169–186.
Abbott, A. (2016). *Processual sociology*. University of Chicago Press.
Akrich, M. (1992). The description of technical objects'. In W. Bijker, & J. Law (Eds.), *Shaping technology/building society*. MIT Press.
Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*. https://doi.org/10.1177/1461444816676645
Bartha, P. (2013). Analogy and analogical reasoning. In *The Stanford encyclopedia of philosophy* (Fall 2013 edition).
Bastani, H., & Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research, 68*, 1.
Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K.,..., & Nagar, S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. Preprint available on: arXiv preprint arXiv:1810.01943.
Boltanski, L. (2011). *On critique. A sociology of emancipation*. Polity.
Boltanski, L., & Thévenot, L. (1983). Finding one's way in social space. *Social Science Information, 22*(4–5), 631–680.
Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of conference on fairness, accountability and transparency, FAT2018* (pp. 77–91).
Burrel, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society, 3*(1), 1–12.
Cai, W., Gaebler, J., Garg, N., & Goel, S. (2020, February). Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 22–28).
Cardon, D., Cointet, J.-P., & Mazières, A. (2018). Neurons spike back. The invention of inductive machines and the artificial intelligence controversy. *Réseaux, 5*(211), 173.
Chatterjee, S., Sarker, S., & Fuller, M. (2009). Ethical information systems development: A Baumanian postmodernist perspective. *Journal of the Association for Information Systems, 10*(11), 3.
Clarke, R. (2019). Principles and business processes for responsible AI. *Computer Law and Security Review*. https://doi.org/10.1016/j.clsr.2019.04.007
Cloutier, C., & Langley, A. (2017). Negotiating the moral aspects of purpose in single and cross-sectoral collaborations. *Journal of Business Ethics, 141*(1), 103–131.
Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*, 311–313.
Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. Retrieved on October 9, 2018.
Datta, A., Datta, A., Makagon, J., Mulligan, D. K., & Tschantz, M. C. (2018). Discrimination in online advertising: A multidisciplinary inquiry. In *Proceedings of the 1st conference on fairness, accountability and transparency, PMLR* (Vol. 81, pp. 20–34).
Desrosières, A. (1998). *The politics of large numbers: A history of statistical reasoning*. Harvard University Press.
Dey, P., & Lehner, O. (2017). Registering ideology in the creation of social entrepreneurs: Intermediary organizations, 'ideal subject' and the promise of enjoyment. *Journal of Business Ethics, 142*(4), 753–767.
Eubanks, V. (2017). *Automating inequality, how high-tech tools profile, police, and punish the poor*. Martin's Press.
Fazelpour, S., & Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. In *AAAI/ACM conference on artificial intelligence, ethics, and society (AIES).*

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy and Technology, 32*, 185–193.

Friedler, S., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. Preprint available on arXiv:1609.07236

Gálvez, A., Tirado, F., & Alcaraz, J. M. (2020). "Oh! Teleworking!" Regimes of engagement and the lived experience of female Spanish teleworkers. *Business Ethics: A European Review, 29*(1), 180–192.

Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted Twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. *Journal of Management Information Systems, 33*(4), 1034–1058.

Greene, D., Hoffmann, A. L., & Stark, L. (2019, January). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Haas, C. (2019). The price of fairness: A framework to explore trade-offs in algorithmic fairness. In *Proceedings of the international conference on information systems (ICIS) 2019*, Munich, December.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523–541.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323).

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms and the limits of antidiscrimination discourse. *Information, Communication and Society, 22*(7), 900–915.

Hoffmaster, B. (2018). From applied ethics to empirical ethics to contextual ethics. *Bioethics*, *32*(2), 119–125.

Hofmann, H. J. (1990). Die Anwendung des CART-Verfahrens zur statistischen Bonitätsanalyse von Konsumentenkrediten. *Zeitschrift fur Betriebswirtschaft, 60*, 941–962.

Hesse, M. (1966). *Models and analogies in science*. University of Notre Dame Press.

Jaton, F. (2021). *The constitution of algorithms: Ground-truthing, programming, formulating*. MIT Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399.

John-Mathews, J. M. (2021). International Conference on Information Systems 2021 Proceedings. 13. https://aisel.aisnet.org/icis2021/ai_business/ai_business/13.

John-Mathews, J. M. (2022). Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technological Forecasting and Social Change*, *174*, 121209.

Kamiran, F., & Calders, T. (2010, May). Classification with no discrimination by preferential sampling. In *Proceedings of the 19th machine learning conference*, *Belgium and The Netherlands* (pp. 1–6).

Kay, M., Matuszek, C., & Munson, S. A. (2015, April). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3819–3828).

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems, 134*, 113302.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*(1), 237–293.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In *Abstracts of the 2018 ACM international conference on measurement and modeling of computer systems (SIGMETRICS '18)*.

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine, 23*(1), 89–109.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science, 65*(7), 2966–2981.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analysed the COMPAS recidivism algorithm. *ProPublica, 5*(2016), 9.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Leclercq-Vandelannoitte, A., & Bertin, E. (2018). From sovereign IT governance to liberal IT governmentality? A Foucauldian analogy. *European Journal of Information Systems, 27*(3), 326–346.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics, 160*(2), 377–392.

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Springer.

Little, J. D. (1970). Models and managers: The concept of a decision calculus. *Management Science, 16*(8), B466.

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair auto encoder. arXiv preprint arXiv:1511.00830.

Lu, T., Zhang, Y., & Li, B. (2019). The value of alternative data in credit risk prediction: Evidence from a large field experiment. In *ICIS 2019 conference*, Munich, December.

Mackenzie, A. (2017). *Machine learners*. The MIT Press.

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 349–358).

Martin, K. (2019a). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*(4), 835–850.

Martin, K. (2019b). Designing ethical algorithms. *MIS Quarterly Executive, 18*(2), 129–142.

Martin, K. E., & Freeman, R. E. (2004). The separation of technology and ethics in business ethics. *Journal of Business Ethics, 53*(4), 353–364.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.

Mercier-Roy, M., & Mailhot, C. (2019). What's in an app? Investigating the moral struggles behind a sharing economy device. *Journal of Business Ethics, 159*(4), 977–996.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 1–7.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society, 3*(2), 2053951716679679.

Molnar, C. (2018). Interpretable machine learning: A guide for making black box model explainable. Retrieved June 6, 2018, from https://christophm.github.io/interpretable-ml-book

Morley, J., Floridi, L., Kinsey, L., et al. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill, 5*(3), e00024-001.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345–1359.

Pentland, A. (2014). *Social physics: How good ideas spread*. Penguin Press.

Powers, T. M., & Ganascia, J. G. (2020). *The ethics of the ethics of AI*. Oxford University Press.

Rouvroy, A., & Berns, T. (2013). Algorithmic governmentality and prospects of emancipation. Disparateness as a precondition for individuation through relationships? *Réseaux, 177*(1), 163–196.

Saha, D., Schumann, C., McElfresh, D. C., Dickerson, J. P., Mazurek, M. L., & Tschantz, M. C. (2020, February). Human comprehension of fairness in machine learning. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 152–152).

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020, January). What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 458–468).

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency FAT2019* (pp. 59–68).

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2239–2248).

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue, 11*(3), 10–29.

Tatman, R. (2016). *Google's speech recognition has a gender bias. Making noise and hearing things*. United Nations.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International conference on machine learning* (pp. 325–333).

Zhou, Y., & Danks, D. (2020, February). Different "Intelligibility" for different folks. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 194–199).

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power, public affairs*. Profile Books.